BIOMETRICS 59, 1027–1035
December 2003

Homogeneity Score Test for the Intraclass Version of the Kappa Statistics and Sample-Size Determination in Multiple or Stratified Studies

Jun-mo Nam

Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS, Executive Plaza South/Room 8028, 6120 Executive Boulevard, MSC 7240, Rockville, Maryland 20852-7240, U.S.A.
email: namj@mail.nih.gov

SUMMARY. When the intraclass correlation coefficient or the equivalent version of the kappa agreement coefficient have been estimated from several independent studies or from a stratified study, we have the problem of comparing the kappa statistics and combining the information regarding the kappa statistics in a common kappa when the assumption of homogeneity of kappa coefficients holds. In this article, using the likelihood score theory extended to nuisance parameters (Tarone, 1988, *Communications in Statistics—Theory and Methods* **17**(5), 1549–1556) we present an efficient homogeneity test for comparing several independent kappa statistics and, also, give a modified homogeneity score method using a noniterative and consistent estimator as an alternative. We provide the sample size using the modified homogeneity score method and compare it with that using the goodness-of-fit method (GOF) (Donner, Eliasziw, and Klar, 1996, *Biometrics* **52,** 176–183). A simulation study for small and moderate sample sizes showed that the actual level of the homogeneity score test using the maximum likelihood estimators (MLEs) of parameters is satisfactorily close to the nominal and it is smaller than those of the modified homogeneity score and the goodness-of-fit tests. We investigated statistical properties of several noniterative estimators of a common kappa. The estimator (Donner et al., 1996) is essentially efficient and can be used as an alternative to the iterative MLE. An efficient interval estimation of a common kappa using the likelihood score method is presented.

KEY WORDS: Estimator of common kappa; Homogeneity tests; Interval estimation; Kappa statistics; Sample size.

1. Introduction

The kappa coefficient has been the widely accepted measure for assessing the degree of agreement between two ratings on the presence or absence of a characteristic, e.g., in biometry, psychometry, and behavior science. There are two types of kappa: Cohen's kappa (1960) and the intraclass version of kappa, which is identical to Scott's index (1955). The former is based on a model that the probability of positive classification by the first rating and that by the second rating are different, while the latter assumes that the two probabilities are the same. In this report, we limit our attention to the intraclass version of the kappa or intra-rater correlation coefficient. Note that the intraclass kappa is algebraically equivalent to the inbreeding coefficient (Wright, 1951) in population genetics. Estimation of the kappa coefficient from a single set of data has been investigated by many authors. For example, point estimation has been investigated by Fleiss and Davies (1982) and Bloch and Kramer (1989), while interval estimation has been investigated by Donner and Eliasziw (1992), Hale and Fleiss (1993), and Nam (2000). When multiple studies or a stratified study have been conducted, we would like to compare kappa statistics and present a common or summary kappa agreement using all available information. Donner and Klar (1996) have discussed the problems of inference on kappa statistics in multiple samples. One of their examples of multiple samples is a meta-analysis using data from six retrospective and four prospective studies related to the accuracy of the technetium bone scanning in the diagnosis of osteomyelitis (Littenberg, Mushlin, and the DTAC, 1992). In the absence of a golden standard for each study, it would be appropriate to examine the studies using the kappa agreement coefficient instead of using sensitivity or specificity. If a homogeneity test for the ten kappa statistics was not rejected, we could make statistical inference on a common kappa summarizing all studies. Otherwise, we could undertake further analysis to determine the source of heterogeneity (between designs, within a design, or both). Another example of the multiple-samples problem is the comparison of twin concordance rates with respect to cigarette-smoking history, by sex and zygosity (Hannah, Hopper, and Mathews, 1985), to investigate if the smoking habit has a genetic component. A chi-square test for homogeneity of kappa coefficients using estimated large-sample variance has been give by Fleiss (1981). Donner et al. (1996) have presented a homogeneity test applying a goodness-of-fit

(GOF) approach and recommended the GOF method over Fleiss's test, for a small sample size based on a result of simulation. Donner (1998) has also provided sample size formula for the comparison of two or more kappas using a GOF test. The GOF procedure is not based on the maximum likelihood estimators of parameters and it may not be fully optimal. In this article, we present the homogeneity score test using MLEs and also a modified homogeneity score test using estimators that are both noniterative and consistent. Our simulations show that the actual level of the score test is closer to nominal in comparison with those of other tests. Problems related to testing homogeneity and summarizing kappa agreement based on multiple samples are in this article. We also investigated estimators of a common kappa and their statistical properties.

In Section 2, we describe the model and notation. In Section 3, we derive the homogeneity score test of several kappa statistics and a modified homogeneity score test using estimators that are simple and consistent. We also examine type 1 error rates of the various homogeneity tests in a simulation study. In Section 4, we compare the sample sizes for a given power using the modified homogeneity score test with that using the GOF procedure. Point estimates of a common kappa are introduced and their variances are derived in Section 5. Using a Monte Carlo experiment several non-iterative estimators are compared in terms of bias and mean square error, for small to moderate sample size and asymptotic relative efficiency, for a large sample size in Section 6. Interval estimation of a common kappa using the score method is presented in Section 7. Section 8 is an example and Section 9 contains concluding remarks.

2. Model and Notation

Consider J independent studies involving n_j subjects for $j = 1, 2, \ldots, J$. Each subject is rated by two examiners, or twice by a given examiner, with ratings denoted as either positive or negative. Denote the probabilities of a positive and a negative rating as $Pr\ (+) = p_j$ and $Pr\ (-) = q_j$, respectively, where $p_j + q_j = 1$ for the jth study. The n_j pairs of ratings can be divided into three categories: (+, +); (+, −) or (−, +); and (−, −). The observed numbers of pairs in the categories are x_{2j}, x_{1j}, and x_{0j} and their corresponding probabilities are P_{2j}, P_{1j}, and P_{0j}, where the first subscript represents the number of positive ratings in a pair for the jth study. Define the kappa, κ_j, as the correlation coefficient between two ratings in a pair, i.e., $\kappa_j = (P_{2j} - p_j^2)/(p_j q_j) = (P_{0j} - q_j^2)/(p_j q_j)$, which yields the following multinomial model: $P_{2j}(\kappa_j, p_j) = p_j^2 + p_j q_j \kappa_j$, $P_{1j}(\kappa_j, p_j) = 2p_j q_j (1 - \kappa_j)$ and $P_{0j}(\kappa_j, p_j) = q_j^2 + p_j q_j \kappa_j$ (Mak, 1988; Bloch and Kramer, 1989) for $j = 1, 2, \ldots, J$. The observed data for the jth study are summarized in Table 1. The intraclass-correlation coefficient, κ_j, is the same as the kappa by the standard definition, i.e., $\kappa_j = (P_{0j} - p_{ej})/(1 - p_{ej})$, where the probability of the observed agreement and that of the expected under independence are $p_{0j} = P_{2j} + P_{0j}$ and $p_{ej} = p_j^2 + q_j^2$, respectively, for every j. Note that the intraclass kappa and correlation coefficient are the same only when the probability of a positive classification is the same for all ratings for the jth study.

Table 1
Observations for the jth study

Category	Observation
(+, +)	x_{2j}
(+, −) or (−, +)	x_{1j}
(−, −)	x_{0j}
Sum	n_j

3. Testing Homogeneity of Kappa Statistics

Consider a test for homogeneity of several kappa statistics, i.e., $H_0 : \kappa_j = \kappa$ for $j = 1, 2, \ldots, J$. From the joint distribution of $x' = (x_1', \ldots, x_J')$ where $x_j' = (x_{2j}, x_{1j}, x_{0j})$ for $j = 1, 2, \ldots, J$, the log likelihood is expressed as $\ln L(\kappa, \mathbf{p}) = \sum_{j=1}^{J} \ln L_j(\kappa_j, p_j)$, where $\kappa' = (\kappa_1, \kappa_2, \ldots, \kappa_J)$ and $\mathbf{p}' = (p_1, p_2, \ldots, p_J)$. The MLEs of κ_j and p_j are $\hat{\kappa}_j = (4x_{2j}x_{0j} - x_{1j}^2)/\{(2x_{2j} + x_{1j})(2x_{0j} + x_{1j})\}$ and $\hat{p}_j = (2x_{2j} + x_{1j})/(2n_j)$, and the variance of $\hat{\kappa}_j$ is $\text{var}(\hat{\kappa}_j) = (1 - \kappa_j)\{(1 - \kappa_j)(1 - 2\kappa_j) + \kappa_j(2 - \kappa_j)/(2p_j q_j)\}/n_j$ for $j = 1, 2, \ldots, J$ (e.g., Bloch and Kramer, 1989). With a common kappa over the J tables, the log likelihood is written as

$$\ln L(\kappa, \boldsymbol{p}) = \sum_{j=1}^{J} \ln L_j(\kappa, p_j), \tag{1}$$

where $\ln L_j(\kappa, p_j) = x_{2j} \cdot \ln\{p_j(p_j + q_j\kappa)\} + x_{1j} \cdot \ln\{2p_j q_j (1 - \kappa)\} + x_{0j} \cdot \ln\{q_j(q_j + p_j\kappa)\}$ and $q_j = 1 - p_j$ for $j = 1, 2, \ldots, J$. Denote partial derivatives as $S_\kappa(\kappa, p_j) \equiv \partial \ln L_j/\partial\kappa$, $S_j(\kappa, p_j) \equiv \partial \ln L_j/\partial p_j$ and $S_\kappa(\kappa, \boldsymbol{p}) = \sum_j S_\kappa(\kappa, p_j)$. The MLEs of κ and $\mathbf{p}$ are the solution of $J + 1$ partial equations, i.e., $S_\kappa(\kappa, \boldsymbol{p}) = 0$ and $S_j(\kappa, p_j) = 0$ for $j = 1, 2, \ldots, J$. They cannot be expressed in a closed form, but can be found numerically by an iterative procedure (Appendix). Note that the MLEs of nuisance parameters, $\tilde{p}_j s$, are not the same as the $\hat{p}_j s$. Define normal deviate as,

$$z_j(\tilde{\kappa}, \tilde{p}_j) = S_\kappa(\tilde{\kappa}, \tilde{p}_j)/\{v(\tilde{\kappa}, \tilde{p}_j)\}^{1/2}$$

where

$$S_\kappa(\tilde{\kappa}, \tilde{p}_j) = \{x_{2j}/(\tilde{p}_j + \tilde{q}_j\tilde{\kappa}) + x_{0j}/(\tilde{q}_j + \tilde{p}_j\tilde{\kappa}) - n_j\}/(1 - \tilde{\kappa}),$$

$$v(\tilde{\kappa}, \tilde{p}_j) = n_j/[(1 - \tilde{\kappa})\{(1 - \tilde{\kappa})(1 - 2\tilde{\kappa}) + \tilde{\kappa}(2 - \tilde{\kappa})/(2\tilde{p}_j\tilde{q}_j)\}],$$

and $\tilde{\kappa}$ and $\tilde{p}_j$ are the MLEs of κ and p_j, and $\tilde{q}_j = 1 - \tilde{p}_j$ for $j = 1, 2, \ldots, J$. The likelihood score statistic for testing $H_0 : \kappa_j = \kappa$ for every j is

$$X_s^2 = \sum_{j=1}^{J} z_j^2(\tilde{\kappa}, \tilde{p}_j), \tag{2}$$

which is asymptotically distributed as a chi-square with J − 1 degrees of freedom. The homogeneity hypothesis is rejected at level α when $X_s^2 \geq \chi^2_{(1-\alpha),J-1}$, where $\chi^2_{(1-\alpha),J-1}$ is the 100x $(1 - \alpha)$ percentile point of the chi-square distribution with J − 1 degrees of freedom. For computational simplicity, in Section 5, we consider several noniterative and consistent estimators of a common kappa. Using the theory of homogeneity score test extended to nuisance parameters (Tarone, 1988), we

find a modified score statistic, e.g.,

$$X_D^2 = \sum_{j=1}^{J} \{S_\kappa(\hat{\kappa}_D, \hat{p}_j)\}^2 / v(\hat{\kappa}_D, \hat{p}_j) - \left\{\sum_{j=1}^{J} S_\kappa(\hat{\kappa}_D, \hat{p}_j)\right\}^2 \Big/ \sum_{j=1}^{J} v(\hat{\kappa}_D, \hat{p}_j), \quad (3)$$

which is asymptotically a chi-square with J − 1 degrees of freedom as $n_j \to \infty$ for a fixed J. The estimator, $\tilde{\kappa}_D$, is defined in (15). If the consistent estimators of κ and p_j's are MLEs, then the second term of (3) vanishes, since $\sum_j S_\kappa(\tilde{\kappa}, \tilde{p}_j) = 0$, and (3) reduces to (2). The chi-square GOF test for homogeneity given by Donner et al. (1996) is

$$X_G^2 = \sum_{i=0}^{2} \sum_{j=1}^{J} \{x_{ij} - n_j P_{ij}(\hat{\kappa}_D, \hat{p}_j)\}^2 / \{n_j P_{ij}(\hat{\kappa}_D, \hat{p}_j)\}. \quad (4)$$

A test similar to Fleiss's is

$$X_F^2 = \sum_{j=1}^{J} \hat{\omega}_j (\hat{\kappa}_j - \hat{\kappa}_\omega)^2, \quad (5)$$

where $\hat{\omega}_j$ is the inverse of an estimator of the variance of $\hat{\kappa}_j$, and $\hat{\kappa}_\omega$ is a weighted average of $\hat{\kappa}_j$'s (see equation [17], Section 5). Under H_0, the GOF statistics (4) and Fleiss's statistics (5) are asymptotically distributed as a chi-square with J − 1 degrees of freedom as $n_j \to \infty$ for a fixed J.

Consider testing homogeneity of kappa statistics assuming $p_j = p$ for $j = 1, 2, \ldots, J$. Under $\kappa_j = \kappa$ and $p_j = p$ for every j, the MLEs of κ and p are $\hat{\kappa} = \hat{\kappa}_p$ (see Section 5) and $\hat{p} = (2x_{2\cdot} + x_{1\cdot})/(2n_\cdot)$, where $x_{i\cdot} = \sum_{j=1}^{J} x_{ij}$ and $n_\cdot = \sum_{j=1}^{J} n_j$. The homogeneity test using the score method is

$$X_{s'}^2 = \left\{1 - 2\hat{\kappa} + \frac{\hat{\kappa}(2-\hat{\kappa})}{2(1-\hat{\kappa})\hat{p}\hat{q}}\right\} \cdot \sum_{j=1}^{J} \left\{\left(\frac{x_{2j}}{\hat{p} + \hat{q}\hat{\kappa}} + \frac{x_{0j}}{\hat{q} + \hat{p}\hat{\kappa}} - n_j\right)^2 \Big/ n_j\right\}. \quad (6)$$

The homogeneity is rejected at level α if $X_{s'}^2 \geq \chi^2_{(1-\alpha),J-1}$. If $\kappa_j = \kappa$ and $p_j = p$ for every j, we can pool the J sets of data for inference on the common kappa.

We investigate type 1 error rates of various homogeneity tests for small or moderate sample sizes. Results of a Monte Carlo experiment with 10,000 simulations for $(p_1, p_2) =$ (0.2, 0.3), (0.2, 0.5), (0.3, 0.5), and $\kappa = 0.2, 0.4, 0.6, 0.8$ for $(n_1, n_2) =$ (20, 30), and (40, 60) are summarized in Table 2. The empirical type 1 error rates of the homogeneity score test using the MLEs were satisfactorily close to a nominal 0.05 level. Those of the GOF and modified score test using $\hat{\kappa}_D$ tended to be anticonservative and greater than those of the score test using MLEs. Fleiss's method provided unreliable type 1 error rates whose discrepancy from the nominal level was excessive: overly anticonservative in general, but conservative when a kappa agreement is very strong. Table 2 indicates that the homogeneity score test using the MLEs was the best testing procedure among four methods in terms of the type 1 error probability for a finite sample. Similar findings were observed using homogeneity tests for kappa statistics from three samples (Table 3) based on a simulation study for $(p_1, p_2, p_3) = (0.1, 0.3, 0.5)$, $(0.3, 0.4, 0.5)$, $\kappa = 0.2, 0.4, 0.6, 0.8$, and $(n_1, n_2, n_3) = (20, 20, 20)$, and $(30, 20, 10)$. The performance of the Fleiss's-type test is clearly unsatisfactory for small and moderate sample sizes. To adjust the anticonservativeness of the GOF test, we may apply the F-distribution in place of the chi-square with J − 1 degrees of freedom, as an approximate distribution of the GOF statistic for small or moderate total sample size. The critical value of the test at α could be $(J-1) \cdot F_{J-1,n.-J+1}(1-\alpha)$, where $F_{J-1,n.-J+1}(1-\alpha)$ is the 100x $(1-\alpha)$ percentile point of the F-distribution with J − 1 and $n \cdot -J + 1$ degrees of freedom. Simulations show that the adjustment can adequately reduce the anticonservativeness of the GOF test.

Table 2

Empirical type 1 error rates of homogeneity tests for $\kappa_1 = \kappa_2 = \kappa$ (10,000 simulations)

		$n_1 = 20, n_2 = 30$				$n_1 = 40, n_2 = 60$			
(p_1, p_2)	κ	X_s^2	X_G^2	X_D^2	X_F^2	X_s^2	X_G^2	X_D^2	X_F^2
(0.2, 0.3)	0.2	.046	.055	.056	.132	.055	.057	.059	.092
	0.4	.056	.064	.067	.133	.052	.055	.059	.078
	0.6	.050	.059	.063	.088	.055	.061	.065	.068
	0.8	.042	.065	.055	.036	.049	.053	.055	.030
(0.2, 0.5)	0.2	.046	.053	.055	.145	.049	.054	.054	.087
	0.4	.050	.055	.058	.131	.053	.057	.060	.079
	0.6	.057	.063	.065	.095	.050	.055	.059	.070
	0.8	.050	.063	.058	.039	.049	.052	.056	.032
(0.3, 0.5)	0.2	.054	.055	.056	.089	.049	.050	.050	.063
	0.4	.055	.058	.059	.083	.055	.056	.057	.063
	0.6	.054	.057	.058	.062	.055	.058	.060	.065
	0.8	.048	.054	.051	.018	.049	.050	.051	.039

Note: X_s^2, X_G^2, X_D^2, and X_F^2 refer to statistics for testing homogeneity of kappas using the score, goodness-of-fit, modified score, and Fleiss methods, respectively.

4. Power and Sample Size

Denote $\bar{\kappa} = \sum_{j=1}^{J} n_j p_j q_j \kappa_j / (\sum_{j=1}^{J} n_j p_j q_j)$. The homogeneity score test using $\hat{\kappa}_D$ (3) under the alternative $\kappa_j \neq \kappa$ is a noncentral chi-square with J − 1 degrees of freedom and noncentrality parameter

$$\lambda_D = \sum n_j c_j d_j^2 - \left(\sum n_j d_j\right)^2 \Big/ \left(\sum n_j / c_j\right) \quad (7)$$

where $c_j = 1 - 2\bar{\kappa} + \bar{\kappa}(2-\bar{\kappa})/\{2(1-\bar{\kappa})p_j q_j\}$ and $d_j = p_j(p_j + q_j\kappa_j)/(p_j + q_j\bar{\kappa}) + q_j(q_j + p_j\kappa_j)/(q_j + p_j\bar{\kappa}) - 1$ for $j = 1, 2, \ldots, J$. The contribution of the second term of (7) is usually negligible. Since the limit value of MLEs, $\tilde{\kappa}$ and $\tilde{p}_j's$, cannot be written in an explicit form, we approximate the noncentrality parameter using $\bar{\kappa}$ and $p's$. The asymptotic power of the modified homogeneity test (6) under the assumption of $p_j = p$ for every j is distributed as a noncentral chi-square with J − 1 degrees of freedom and noncentrality parameter

$$\lambda'_D = c\left\{\frac{(1+\bar{\kappa})pq}{\bar{\kappa} + (1-\bar{\kappa})^2 pq}\right\}^2 \cdot \left\{\sum_{j=1}^{J} n_j (\kappa_j - \bar{\kappa})^2\right\} \quad (8)$$

Table 3
Empirical Type 1 error rates of homogeneity tests for $\kappa_1 = \kappa_2 = \kappa_3 = \kappa$ (10,000 simulations)

		$n_1 = n_2 = n_3 = 20$				$n_1 = 30, n_2 = 20, n_3 = 10$			
(p_1, p_2, p_3)	κ	X_s^2	X_G^2	X_D^2	X_F^2	X_s^2	X_G^2	X_D^2	X_F^2
(0.1, 0.3, 0.5)	0.2	.036	.054	.055	.232	.031	.065	.063	.206
	0.4	.044	.047	.053	.326	.042	.058	.061	.235
	0.6	0.44	.054	.060	.299	.044	.058	.063	.227
	0.8	.053	.070	.071	.209	.046	.068	.064	.148
(0.3, 0.4, 0.5)	0.2	.046	.053	.053	.111	.040	.055	.055	.131
	0.4	.049	.057	.058	.103	.045	.052	.053	.121
	0.6	.046	.055	.056	.080	.048	.058	.060	.158
	0.8	.037	.045	.045	.047	.040	.049	.049	.140

Note: X_s^2, X_G^2, X_D^2, and X_F^2 refer to statistics for testing homogeneity of kappas using the score, goodness-of-fit, modified score, and Fleiss methods, respectively.

where $c = 1 - 2\bar{\kappa} + \bar{\kappa}(2 - \bar{\kappa})/\{2(1 - \bar{\kappa})pq\}$. The power of the test can be obtained from tables of the cumulative noncentral chi-square distribution (Haynam, Govindarajulu, and Leone, 1970).

Define design parameters as $t_j = n_j/n.$ for $j = 1, 2, \ldots, J$. From the noncentrality parameter λ_D, the total sample size required for power $1 - \beta$ of the modified homogeneity score test at level α can be found by

$$n. = \lambda(J-1, 1-\beta, \alpha) \Bigg/ \left\{ \sum_{j=1}^{J} t_j c_j d_j^2 - \left(\sum_{j=1}^{J} t_j d_j \right)^2 \Bigg/ \left(\sum_{j=1}^{J} t_j / c_j \right) \right\}, \quad (9)$$

where $\lambda(J - 1, 1 - \beta, \alpha)$ is the value of the noncentrality parameter of the cumulative chi-square distribution corresponding to power $1 - \beta$ and level α, e.g., $\lambda\,(1, 0.8, 0.05) = 7.849$ for $J = 2$, 80% power and 5% level. The approximate sample size for the jth group is $n_j = t_j \cdot n.$. Similarly, we can find the sample size for a given power of the test (6) from λ'_D (8). Under $p_j = p$ for every j, the second terms of (7) and (9) are zero.

The GOF test for homogeneity under the alternative is asymptotically a noncentral chi-square with J − 1 degrees of freedom and noncentrality parameter

$$\lambda_G = \sum_{i=0}^{2} \sum_{j=1}^{J} n_j \{P_{ij}(\kappa_j, p_j) - P_{ij}(\bar{\kappa}, p_j)\}^2 / P_{ij}(\bar{\kappa}, p_j) \quad (10)$$

(e.g., Donner, 1998), which under the assumption that $p_j = p$ for every j reduces to

$$\lambda'_G = \frac{pq}{(1-\bar{\kappa})} \left\{ 1 + \frac{1}{\bar{\kappa} + (1-\bar{\kappa})^2} \right\} \cdot \left\{ \sum_{j=1}^{J} n_j (\kappa_j - \bar{\kappa})^2 \right\}. \quad (11)$$

From λ_G, we have the total sample size required for a study as

$$n. = \lambda(J-1, 1-\beta, \alpha) \Bigg/ \left[\sum_{i=0}^{2} \sum_{j=1}^{J} t_j \{P_{ij}(\kappa_j, p_j) - P_{ij}(\bar{\kappa}, p_j)\}^2 / P_{ij}(\bar{\kappa}, p_j) \right]. \quad (12)$$

For the special case that $n_j = n$ and under $p_j = p$ for all j, we have

$$n = \lambda(J-1, 1-\beta, \alpha) \cdot (1-\bar{\kappa}) \Bigg/ \left[pq \left\{ 1 + \frac{1}{\bar{\kappa} + (1-\bar{\kappa})^2} \right\} \cdot \sum (\kappa_j - \bar{\kappa})^2 \right].$$

Using (9) and (12), approximate sample sizes required for 80% power of the score and GOF tests using $\hat{\kappa}_D$ for detecting heterogeneity of two kappa statistics at $\alpha = 0.05$ for various values of the prevalence rate, kappa coefficient and design parameters are summarized in Table 4. The sample size required for a given power of the score test is always smaller than that of the GOF test, except when $p_1 = p_2 = 0.5$, in which case they are equal. A very large sample size is needed when the difference between two kappas is small and/or $|p's - 0.5|$ is large. Numerical investigation shows that the balanced design, $t_1 = t_2 = 0.5$, is most efficient. Similar observations are found for sample size requirement for 80% power of the homogeneity tests at $\alpha = 0.05$ for $J = 3$. A simulation study was performed to examine whether the asymptotic sample size formula (9) and (12) can be satisfactorily applied in finite samples. Table 4 shows that the actual power of the homogeneity test for given approximate sample size was reasonably close to the nominal power for both the modified score and the GOF tests. In particular, sample sizes calculated under a balanced design ($t_1 = t_2$) provided the intended power on average. Those under an unbalanced design tend to be either slightly underestimated or overestimated, depending on ($t_1 < t_2$) or ($t_1 > t_2$). The actual levels of both the modified score and the GOF tests are anticonservative. The former is a little more conservative than the latter.

5. Estimation of Common Kappa

When several kappa statistics from independent studies or a stratified study are given, we would like to examine the homogeneity of kappas. An estimate of the common kappa is required for constructing a homogeneity test. If the homogeneity assumption is reasonable, the estimate of the common kappa can be used as appropriating summary measure reliability.

Table 4

Sample-size requirements for 80% power of homogeneity tests for kappa statistics at α .05 for $J = 2$ (numbers in parentheses are empirical powers)

				$t_1 = t_2 = 0.5$		$t_1 = 0.25, t_2 = 0.75$		$t_1 = 0.75, t_2 = 0.25$	
p_1	p_2	κ_1	κ_2	Modified score X_D^2	GOF X_G^2	Modified score X_D^2	GOF X_G^2	Modified score X_D^2	GOF X_G^2
0.1	0.3	0.2	0.4	1,153 (.80)	1,207 (.80)	1,774 (.81)	1,882 (.81)	1,308 (.78)	1,342 (.79)
			0.6	266 (.80)	281 (.80)	386 (.77)	413 (.78)	317 (.79)	329 (.82)
			0.8	98 (.80)	104 (.82)	122 (.78)	129 (.78)	132 (.83)	138 (.83)
		0.4	0.6	995 (.79)	1,054 (.79)	1,494 (.78)	1,597 (.80)	1,159 (.79)	1,210 (.80)
			0.8	195 (.79)	206 (.79)	256 (.76)	269 (.75)	259 (.83)	270 (.83)
		0.6	0.8	665 (.77)	697 (.79)	939 (.75)	986 (.76)	832 (.82)	865 (.79)
0.2	0.2	0.2	0.4	970 (.79)	1,000 (.81)	1,272 (.78)	1,315 (.80)	1,302 (.79)	1,334 (.79)
			0.6	232 (.82)	241 (.78)	283 (.79)	294 (.81)	323 (.80)	333 (.78)
			0.8	94 (.83)	98 (.83)	99 (.77)	103 (.78)	141 (.83)	146 (.84)
		0.4	0.6	848 (.80)	883 (.78)	1,061 (.76)	1,106 (.78)	1,189 (.81)	1,237 (.80)
			0.8	184 (.81)	192 (.81)	199 (.81)	206 (.78)	283 (.81)	294 (.82)
		0.6	0.8	596 (.79)	618 (.79)	687 (.77)	710 (.77)	894 (.80)	929 (.83)
Average power				.80	.80	.78	.78	.81	.81

Consider the estimation of a common kappa, κ, over J strata or J sets of data. The MLE of κ based on the pooled data is $\hat{\kappa}_p = (4x_{2.}x_{0.} - x_{1.}^2)/\{(2x_{2.} + x_{1.})(2x_{0.} + x_{1.})\}$, where summation is denoted by dots, e.g., $x_{i.} = \sum_{j=1}^{J} x_j$ for $i = 0$, 1, and 2. The pooled kappa is not a consistent estimator, unless $p_j = p$ for every j. A chi-square GOF test for $\kappa_j = \kappa$ and $p_j = p$ for $j = 1, 2, \ldots, J$ can be used to determine whether pooling is appropriate.

The MLEs of a common kappa and the nuisance parameters, $\tilde{\kappa}$ and $\tilde{p}_j$s, cannot be expressed in a closed form. They can be obtained by an iterative procedure (see the Appendix). From the inversion of the information matrix, we obtain the variance of the MLE of κ as

$$\text{var}(\tilde{\kappa}) = (1-\kappa) \Bigg/ \left[\sum_{j=1}^{J} \frac{n_j}{\{(1-\kappa)(1-2\kappa) + \kappa(2-\kappa)/(2p_jq_j)\}} \right]. \tag{13}$$

The asymptotic variance of the MLE of κ for a single set (e.g., Bloch and Kramer, 1989) is a special case of (13) for $J = 1$.

Consider noniterative estimators. Similar in form to the Mentel-Haenszel estimator, the ratio of the sum of numerators and the sum of denominators from individual $\hat{\kappa}_j$'s, $\hat{\kappa}_{\text{MH}} = \sum_j (4x_{2j}x_{0j} - x_{1j}^2) / \sum_j (2x_{2j} + x_{1j})(2x_{0j} + x_{1j})$ is consistent. Letting $b_1 = 2(1-\kappa)(1-2\kappa)$ and $b_2 = \kappa(2-\kappa)$, the asymptotic variance is found as

$$\text{var}(\hat{\kappa}_{\text{MH}}) = (1-\kappa) \left\{ b_1 \left(\sum_j n_j^3 p_j^2 q_j^2 \right) + b_2 \left(\sum_j n_j^3 p_j q_j \right) \right\} \Bigg/ \left\{ 2 \left(\sum n_j^2 p_j q_j \right)^2 \right\}. \tag{14}$$

Donner et al. (1996) suggested an estimator of κ as

$$\hat{\kappa}_D = \sum_j n_j \hat{p}_j \hat{q}_j \hat{\kappa}_j \Bigg/ \left(\sum_j n_j \hat{p}_j \hat{q}_j \right). \tag{15}$$

The variance is approximated as

$$\text{var}(\hat{\kappa}_D) = (1-\kappa) \left\{ b_2 \left(\sum_j n_j p_j q_j \right) + b_1 \left(\sum_j n_j p_j^2 q_j^2 \right) \right\} \Bigg/ \left\{ 2 \left(\sum_j n_j p_j q_j \right)^2 \right\}. \tag{16}$$

When $n_j = n$ for every j, the $\hat{\kappa}_{\text{MH}}$ and $\hat{\kappa}_D$ are identical. Denoting $\omega_j = \{\text{var}(\hat{\kappa}_j)\}^{-1}$, a weighted average of individual kappa is written as

$$\hat{\kappa}_\omega = \sum_j \hat{\omega}_j \hat{\kappa}_j \Bigg/ \left(\sum_j \hat{\omega}_j \right) \text{ where } \hat{\omega}_j = \{\text{var}(\hat{\kappa}_j)\}^{-1}_{\kappa=\hat{\kappa}_j, p_j=\hat{p}_j}, \tag{17}$$

(e.g., Fleiss, 1981). An approximated variance of $\hat{\kappa}_\omega$ is expressed as $\text{var}(\hat{\kappa}_\omega) \approx (\Sigma\omega_j)^{-1}$ (e.g., Fleiss and Davies, 1982) which is smaller than the true variance. The $\hat{\kappa}_D$ and $\hat{\kappa}_\omega$ are both consistent. The weighted kappa is undefined whenever any $\hat{p}_j\hat{q}_j$ is zero, while the $\hat{\kappa}_{\text{MH}}$ and $\hat{\kappa}_D$ are defined, unless all $\hat{p}_j\hat{q}_j$'s are zero.

6. Numerical Evaluation on Noniterative Estimators

To evaluate the bias and mean square errors of the noniterative estimators for small or medium sample sizes, a simulation study for the case of two strata was performed under the following configuration: $(p_1, p_2) = (0.1, 0.5), (0.2, 0.5)$, $(n_1, n_2) = (20, 30), (40, 60)$, $\kappa = 0.1, 0.3, 0.5, 0.8$. Results are summarized in Table 5. The bias of $\hat{\kappa}_p$ was not reduced by increased sample size when the p's were different, demonstrating the inconsistency of the pooled estimators. The absolute bias and mean square error of the weighted estimator were far greater than those of $\hat{\kappa}_{\text{MH}}$ and $\hat{\kappa}_D$. Since $\hat{\omega}_j$ and $\hat{\kappa}_j$ are correlated, the $\hat{\kappa}_\omega$ had relatively large bias and standard error, particularly for a small sample size. The estimators, $\hat{\kappa}_{\text{MH}}$ and $\hat{\kappa}_D$, had smaller absolute bias and mean square error. As n_j's increase, the bias and the mean square error of each estimator approached zero and the square root of the asymptotic

Table 5
Bias and mean square error based on 10,000 simulations

(p_1, p_2)	κ	$\hat{\kappa}_p$ bias (MSE$^{1/2}$)	$\hat{\kappa}_{MH}$ bias (MSE$^{1/2}$)	$\hat{\kappa}_D$ bias (MSE$^{1/2}$)	$\hat{\kappa}_\omega$ bias (MSE$^{1/2}$)
$n_1 = 20, n_2 = 30$					
(0.1, 0.5)	0.1	.146 (.205)	−.018 (.167)	−.019 (.162)	−.103 (.178)
	0.3	.113 (.177)	−.016 (.160)	−.017 (.158)	−.147 (.266)
	0.5	.080 (.146)	−.013 (.147)	−.014 (.146)	−.160 (.319)
	0.8	.031 (.090)	−.007 (.103)	−.008 (.103)	−.164 (.348)
(0.2, 0.5)	0.1	.074 (.161)	−.019 (.156)	−.020 (.152)	−.065 (.186)
	0.3	.057 (.148)	−.017 (.150)	−.018 (.148)	−.053 (.206)
	0.5	.041 (.130)	−.013 (.138)	−.014 (.136)	−.031 (.198)
	0.8	.015 (.087)	−.007 (.097)	−.008 (.096)	−.033 (.154)
$n_1 = 40, n_2 = 60$					
(0.1, 0.5)	0.1	.150 (.181)	−.008 (.115)	−.009 (.111)	−.070 (.145)
	0.3	.115 (.150)	−.010 (.114)	−.010 (.112)	−.070 (.196)
	0.5	.083 (.120)	−.007 (.104)	−.007 (.103)	−.045 (.195)
	0.8	.033 (.067)	−.004 (.071)	−.004 (.070)	−.026 (.159)
(0.2, 0.5)	0.1	.078 (.127)	−.009 (.107)	−.010 (.104)	−.029 (.122)
	0.3	.059 (.113)	−.010 (.107)	−.010 (.105)	−.016 (.123)
	0.5	.043 (.097)	−.007 (.098)	−.008 (.096)	−.000 (.110)
	0.8	.017 (.061)	−.004 (.066)	−.004 (.066)	.007 (.070)

variance, respectively. The mean square error of $\hat{\kappa}_D$ tended to be slightly smaller than that of $\hat{\kappa}_{MH}$. Results for $(n_1, n_2) = (10, 40)$ and $(20, 80)$ lead to similar conclusions.

The asymptotic standard errors (the square root of the asymptotic variance) of $\hat{\kappa}_{MH}$, $\hat{\kappa}_D$ and $\tilde{\kappa}$ were calculated for various values of parameters and sample size from (14), (16), and (13) and summarized in Table 6. The estimators, $\hat{\kappa}_{MH}$ and $\hat{\kappa}_D$ possess high relative efficiency, close to 100%. In particular, $\hat{\kappa}_D$ is virtually fully efficient.

7. Interval Estimation of Common Kappa

Consider interval estimation of intraclass kappa coefficient across strata. The likelihood score and its variance are $S_\kappa(\kappa, \boldsymbol{p}) = \sum_j S_\kappa(\kappa, p_j)$ and $\text{var}\{S_\kappa(\kappa, \boldsymbol{p})\} = \sum_j \text{var}\{S_\kappa(\kappa, p_j)\}$, where $S_\kappa(\kappa, p_j) = \{x_{2j}/(p_j + q_j\kappa) + x_{0j}/(q_j + p_j\kappa) - n_j\}/(1-\kappa)$ and $\text{var}\{S_\kappa(\kappa, p_j)\} = n_j/[(1-\kappa)\{(1-\kappa)(1-2\kappa) + \kappa(2-\kappa)/(2p_jq_j)\}]$. As $n_j \to \infty$ for $j = 1, 2, \ldots, J$, $z_s(\kappa, \tilde{\boldsymbol{p}}') \equiv S_\kappa(\kappa, \tilde{\boldsymbol{p}}')/[\text{var}\{S_\kappa(\kappa, \tilde{\boldsymbol{p}}')\}]^{1/2}_{p=\tilde{p}}$ is asymptotically normal with mean zero and variance one. For each $j, \tilde{\boldsymbol{p}}'$ is the MLE of p_j for a given value of κ and a solution of $\partial \ln L_j/\partial p_j = 0$, which is a root of a cubic equation, $a_{oj}\tilde{p}_j'^3 + a_{1j}\tilde{p}_j'^2 + a_{2j}\tilde{p}_j' + a_{3j} = 0$ where $a_{oj} = 2n_j(1-\kappa)^2, a_{1j} = -\{3n_j(1-\kappa) + x_{2j} - x_{0j}\}(1-\kappa)$, $a_{2j} = 2x_{2j} + x_{1j} - 2(2n_j - x_{0j})\kappa + n_j\kappa^2$ and $a_{3j} = (x_{1j} + x_{2j})\kappa$ for $j = 1, 2, \ldots, J$. The approximate $1-\alpha$ confidence limits of κ using the score method is found by solving

$$z_s^2(\kappa, \tilde{\boldsymbol{p}}') = z^2_{(\alpha/2)} \tag{18}$$

Table 6
Standard errors of $\hat{\kappa}_{MH}$, $\hat{\kappa}_D$, and $\hat{\kappa}_{ML}$

(p_1, p_2)	κ	SE($\hat{\kappa}_{MH}$)	SE($\hat{\kappa}_D$)	SE($\hat{\kappa}_{ML}$)	SE($\hat{\kappa}_{MH}$)	SE($\hat{\kappa}_D$)	SE($\hat{\kappa}_{ML}$)
		$n_1 = 20, n_2 = 30$			$n_1 = 10, n_2 = 40$		
(0.1, 0.5)	0.1	.161	.156	.153	.154	.148	.146
	0.3	.157	.154	.154	.148	.144	.144
	0.5	.143	.142	.142	.134	.131	.131
	0.8	.100	.099	.099	.093	.091	.091
(0.2, 0.5)	0.1	.151	.147	.146	.152	.144	.143
	0.3	.147	.144	.144	.146	.139	.139
	0.5	.134	.132	.132	.132	.127	.127
	0.8	.093	.092	.092	.092	.088	.088
		$n_1 = 40, n_2 = 60$			$n_1 = 20, n_2 = 80$		
(0.1, 0.5)	0.1	.114	.111	.108	.109	.105	.104
	0.3	.111	.109	.109	.105	.102	.101
	0.5	.101	.100	.100	.095	.093	.093
	0.8	.070	.070	.070	.066	.064	.064
(0.2, 0.5)	0.1	.107	.104	.103.	.107	.102	.101
	0.3	.104	.102	.102	.103	.098	.098
	0.5	.095	.094	.094	.094	.090	.090
	0.8	.066	.065	.065	.065	.062	.062

where $\tilde{p}'$ is a vector of MLEs of p for a given value of κ. There is only one relevant root, $\tilde{p}'_j = -2(-c_{1j}/3)^{1/2} \cdot \cos(\pi/3 + \theta_j/3) - b_{1j}/3$ where $\cos\theta_j = (27)^{1/2} \cdot c_{2j}/\{2c_{1j}(1-c_{1j})^{1/2}\}$ with $b_{ij} = a_{ij}/a_{0j}$ for i = 1, 2, 3 and $c_{2j} = b_{3j} - b_{1j}b_{2j}/3 + 2(b_{1j}/3)^3$ for $j = 1, 2, \ldots, J$. The lower and upper limits are found by an iterative procedure. This is an extension of the method of interval estimation of the kappa coefficient using the likelihood score method (Nam, 2000). From Section 5, we have simple confidence intervals for κ estimates as $\hat{\kappa}_{\text{MH}} \pm z_{(\alpha/2)} \cdot \{\hat{\text{var}}(\hat{\kappa}_{\text{MH}})\}^{1/2}, \hat{\kappa}_D \pm z_{(\alpha/2)} \cdot \{\hat{\text{var}}(\hat{\kappa}_D)\}^{1/2}$ and $\hat{\kappa}_\omega \pm z_{(\alpha/2)} \cdot \{\hat{\text{var}}(\hat{\kappa}_\omega)\}^{1/2}$. These methods are noniterative and computationally simple. However, for a small sample size and strong kappa agreement, the upper limit of a simple $(1 - \alpha)$ confidence interval for κ may be greater than one. To remedy an unacceptable upper limit, we may use a transformation of $\hat{\kappa}$ similar to Fisher's z-transformation of a sample correlation coefficient (Fisher, 1921), e.g., $g(\hat{\kappa}) = \ln\{(1+\hat{\kappa})/(1-\hat{\kappa})\}$ where $\hat{\kappa}$ may be $\hat{\kappa}_{\text{MH}}$, $\hat{\kappa}_D$, $\hat{\kappa}_\omega$ or $\tilde{\kappa}$. The asymptotic variance of $g(\hat{\kappa})$ is $\text{var}\{g(\hat{\kappa})\} = 4 \cdot \text{var}(\hat{\kappa})/(1-\kappa^2)^2$ using the delta method. A simple $(1 - \alpha)$ confidence interval is given as $g(\hat{\kappa}) \pm z_{(\alpha/2)} \cdot [\hat{\text{var}}\{g(\hat{\kappa})\}]^{1/2}$. By transforming endpoints of this confidence interval back to the original scale of $\hat{\kappa}$, we have adjusted confidence limits that are within $[-1, 1]$. The score method (18) always provides a permissible interval and a reliable confidence coefficient.

As an example, consider two sets of data: $x_{21} = 2$, $x_{11} = 1, x_{01} = 17$ for the first set and $x_{22} = 5, x_{12} = 1$ and $x_{02} = 8$ for the second set, the estimated marginal probabilities of a positive rating and the kappas are $\hat{p}_1 = 0.125$, $\hat{p}_2 = 0.393, \hat{\kappa}_1 = 0.771$ and $\hat{\kappa}_2 = 0.850$. The homogeneity of the two kappa coefficients is not rejected by the homogeneity tests (p values ≈ 0.75). Crude 95% confidence intervals for a common kappa using $\hat{\kappa}_{\text{MH}}, \hat{\kappa}_D, \hat{\kappa}_\omega$ and $\tilde{\kappa}$ with corresponding standard errors are (0.569, 1.055), (0.578, 1.060), (0.590, 1.066), and (0.573, 1.059), respectively. The various confidence intervals are somewhat similar. The upper limits are not acceptable, because they exceed one. Applying Fisher's z-like transformation, the corresponding adjusted simple confidence intervals are (0.382, 0.953), (0.398, 0.955), (0.404, 0.958), and (0.400, 0.953). The upper bounds are less than one. Using the score method (18), we obtain the admissible 95% confidence interval as (0.462, 0.951). Note that the length of the confidence interval by the efficient score method is smaller than those of adjusted simple intervals. The 95% confidence interval found by the GOF procedure (Donner and Eliasziw, 1992) or the score method (Nam, 2000) on pooled data may be used as a trial interval.

8. An Example

Twins on Drinking Status: Equal numbers of male twins and female twins, and equal numbers of monozygote (MZ) and dizygote (DZ) pairs in Melbourne, Australia, were randomly drawn from the Australian National Health and Medical Council Twin Registry for a voluntary interview. Seventy-five pairs were invited for an interview from each sex by zygosity. The drinking information from 181 twin pairs is summarized in Table 7 (Hannah, Hopper, and Mathews, 1983).

For males, the four different homogeneity tests (2), (3), (4), and (5), each led to a p value = 0.02: the twin correlation of MZ twins ($\hat{\kappa}_1 = 0.462$) was significantly larger than that of DZ twins ($\hat{\kappa}_2 = -0.033$) in drinking status. For females, the four homogeneity tests gave p values ($0.55 \sim 0.56$): the kappa agreement of MZ and that of DZ twins, $\hat{\kappa}_1 = 0.474$ and $\hat{\kappa}_2 = 0.360$, were not significantly different. Estimates of a common kappa for females were $\hat{\kappa}_D = 0.414$, $\hat{\kappa}_\omega = 0.417$, and $\tilde{\kappa} = 0.413$. The $\hat{\kappa}_D$ estimate was almost the same as the MLE $\tilde{\kappa}$. Note that the second term of (3) was zero for both male and female twins when $\hat{\kappa}_D$ was used. The homogeneity score test using a simple estimate ($\hat{\kappa}_D$) was similar to that using the MLE of a common kappa in this example. The assumption of $p_1 = p_2$ was quite reasonable for both male and females. Using (6) under $p_1 = p_2$, the homogeneity test for kappa's of MZ and DZ twins for males provided p value = 0.023 and for females gave p value = 0.58. They were slightly larger than those values obtained without the assumption of equal prevalence. For males, the kappa agreement on alcohol use for MZ twin pairs was significantly greater than that of DZ twins. For females, although the kappa agreement for MZ twins was not significantly different from that for DZ twins, the former was larger than the latter. Hannah et al. (1983) suggested that the MZ twins were more concordant than the DZ twins on alcohol use.

9. Discussion

In assessing the degree of agreement in a reliability study, researchers may have a specific value of kappa in mind. Nam (2002) presented an efficient statistic for testing the strength of kappa agreement using the likelihood score and derived a sample size formula for designing a study. In the current article, we investigate statistical methods

Table 7

Like-sex twin pairs by sex, zygosity, and drinking status

	Male			Female		
Alcohol drinking	MZ	DZ	Total	MZ	DZ	Total
Both	19	8	27	11	10	21
One	14	16	30	11	15	26
Neither	19	7	26	23	28	51
Total	52	31	83	45	53	98
$\hat{p}_i$	0.500	0.516		0.367	0.330	
$\hat{\kappa}_i \pm$ SE	0.462 ± 0.123	-0.033 ± 0.179		0.474 ± 0.136	0.360 ± 0.135	

involving a comparison of several kappa statistics, sample size requirement for a set of multiple reliability studies, and estimation of a common kappa using all available information.

Power is highly correlated with the type 1 error probability. Since the actual type 1 error rates for a nominal level of the homogeneity tests are not the same for small or moderate sample sizes, a comparison of tests using empirical power is not straightforward. The power of an anticonservative test is inflated, while that of a conservative test is deflated. Comparison would require careful adjustment of the nominal level of each test so that the actual levels of the tests become the same for a given configuration; then, using the adjusted nominal level, we could generate the power and compare them with equal type 1 error rates. Using repeated trials, for example, we obtain adjusted critical values of the score, GOF and modified score tests at an empirical level of 0.05 as 3.40, 3.74, and 3.86 for given $(p_1, p_2) = (0.2, 0.3)$ and $(n_1, n_2) = (20, 30)$. Empirical powers of the score, GOF and modified score tests for $\kappa_1 = 0.1$ and $\kappa_2 = (0.1, 0.3, 0.5, 0.7, 0.9)$ are (.050, .101, .289, .575, .906), (.050, .100, .294, .565, .895), and (.050, .101, .286, .566, .894), respectively. The three tests are comparable in power, with a slight advantage by the score test.

Since the kappa statistic depends on the prevalence or base rate, some authors, e.g., Thompson and Walters (1988), warned against a comparison of kappa statistics when the prevalence of several studies are different. Others, e.g., Donner et al. (1996), suggested that a comparison of kappa statistics under no assumption of equal prevalence could provide a meaningful assessment on levels of interobserver agreement when the difference among prevalence was not a major concern. They called this a "pragmatic approach." They also noted the difficulty involving validation of equal prevalence in a typical reliability study.

When homogeneity of kappa statistics is not rejected, we may have an interest in a confidence interval for the common kappa, for the evaluation of the kappa agreement coefficient. An efficient interval estimation of the kappa coefficient is derived by extending the likelihood score method (Nam, 2000). The GOF procedure can be also used to derive a relevant confidence interval.

When homogeneity is rejected, we may proceed with a further analysis to investigate the source of heterogeneity. We may partition heterogeneity chi-square approximately into orthogonal components related to specific sources of variation, and examine the significance for each component (e.g., Donner and Klar, 1996). We can investigate a similar approach using the heterogeneity score method.

If both assumptions of homogeneity of kappa's and equal prevalence across strata are valid, then we could perform statistical analysis on the pooled data. However, if the equality of prevalence is not valid, the pooled estimator of a common kappa is not consistent. We should be cautious when pooling data for the assessment of a common kappa in this situation. The homogeneity test based on the GOF procedure and the modified score test are generally close to the homogeneity score method using MLEs of parameters, which is theoretically optimum.

Acknowledgements

The author is very grateful to two referees and an associate editor for their helpful suggestions and constructive comments on an earlier version of the article.

Résumé

Quand le coefficient de corrélation intra-classe ou sa version équivalente du coefficient d'agrément kappa ont été estimés à partir de plusieurs études indépendantes ou d'une étude avec stratification, nous nous trouvons devant le problème de la comparaison de statistiques kappa et de la combinaison d'information sur ces statistiques en un kappa commun quand leur supposition d'homogénéité est satisfaite. Dans cet article, en utilisant la théorie de vraisemblance du score étendue aux paramètres de nuisance (Tarone, 1988) nous présentons un test d'homogénéité efficace pour comparer plusieurs statistiques kappa indépendantes et, de surcroît, nous donnons une méthode du score d'homogénéité modifiée en employant comme alternative un estimateur non-itératif et consistant. Nous fournissons la taille d'échantillon en employant la méthode du score d'homogénéité modifiée et nous la comparons à celle de la qualité d'ajustement (GOF) (Donner, Eliasziw et Klar, 1996). Une étude de simulation pour des tailles d'échantillon faibles et modérés a montré que le niveau actuel du test du score d'homogénéité employant les estimateurs du maximum de vraisemblance (MLEs) des paramètres est, de manière satisfaisante, proche du nominal et qu'il est plus faible que ceux du score d'homogénéité modifiée et des tests de la qualité d'ajustement. Nous étudions les propriétés statistiques de plusieurs estimateurs non itératifs d'un kappa commun. L'estimateur (Donner, Eliasziw et Klar, 1996) est essentiellement efficace et peut être employé comme une alternative au MLE itératif. Nous présentons une estimation par intervalle efficace d'un kappa commun en employant la méthode de vraisemblance du score.

References

Bloch, D. A. and Kramer, H. C. (1989). 2 × 2 kappa coefficients: Measure of agreement or association. *Biometrics* **45,** 269–287.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20,** 37–40.

Donner, A. (1998). Sample size requirements for the comparison of two or more coefficients of inter-observer agreement. *Statistics in Medicine* **17,** 1157–1168.

Donner, A. and Eliasziw, M. (1992). A goodness-of-fit approach to inference procedures for the kappa statistic: Confidence interval consideration, significance-testing and sample-size estimation. *Statistics in Medicine* **11,** 1511–1519.

Donner, A. and Klar, N. (1996). The statistical analysis of kappa in multiple samples. *Journal of Clinical Epidemiology* **49,** 1053–1058.

Donner, A., Eliasziw, M., and Klar, N. (1996). Testing the homogeneity of kappa statistic. *Biometrics* **52,** 176–183.

Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron* **1,** 1–32.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions,* 2nd edition. New York: Wiley.

Fleiss, J. L. and Davies, M. (1982). Jackknifing functions of multifunctional frequencies, with an application to a measure of concordance. *American Journal of Epidemiology* **115,** 841–845.

Hale, C. A. and Fleiss, J. L. (1993). Interval estimation under two study designs for kappa with binary classifications. *Biometrics* **49,** 523–524.

Hannah, M. C., Hopper, J. L., and Mathews, J. D. (1983). Twin concordance for a binary trait: I. Statistical models illustrated with data on drinking status. *Acta Geneticae Medicae et Gemellologiae* **32,** 127–137.

Hannah, M. C., Hopper, J. L., and Mathews, J. D. (1985). Twin concordance for a binary trait: II. Nested analysis of ever-smoking and ex-smoking traits and unnested analysis of a "committed-smoking" trait. *American Journal of Human Genetics* **37,** 153–165.

Haynam, G. E., Govindarajulu, Z., and Leone, G. C. (1970). *Tables of the cumulative non-central chi-square distribution.* Case Statistical Laboratory Publication 104. [Parts of the tables have been published in *Selected Tables in Mathematical Statistics,* Volume 1, H. L. Harter and D. B. Owen (eds).]

Littenberg, B., Mushlin, A. I., and the Diagnostic Technology Assessment Consortium (1992). Technetium bone scanning in the diagnosis of osteomyelitis: A meta-analysis of test performance. *Journal of General Internal Medicine* **7,** 158–163.

Mak, T. K. (1988). Analyzing intraclass correlation for dichotomous variables. *Applied Statistics* **37,** 344–352.

Nam, J. (2000). Interval estimation of the kappa coefficient with binary classification and an equal marginal probability model. *Biometrics* **56,** 583–585.

Nam, J. (2002). Testing the intraclass version of kappa coefficient of agreement with binary scale and sample size determination. *Biometrical Journal* **44,** 558–570.

Scott, W. A. (1955). Reliability of content analysis: The case of normal scale coding. *Public Opinion Quarterly* **19,** 321–325.

Tarone, R. E. (1988). Homogeneity score tests with nuisance parameters. *Communications in Statistics—Theory and Methods* **17**(5), 1549–1556.

Thompson, W. D. and Walters, S. D. (1988). A reappraisal of the kappa coefficient. *Journal of Clinical Epidemiology* **51,** 949–958.

Wright, S. (1951). The general structure of populations. *Annals of Engenics* **15,** 322–354.

Received March 2002. *Revised April* 2003.
Accepted April 2003.

APPENDIX

Maximum Likelihood Estimators of a Common Kappa and Nuisance Parameters

From the second partials of (1), we have the $(J+1) \times (J+1)$ information matrix $\mathbf{I}$ whose elements are

$$I_{00} = -E\left(\frac{\partial^2 \ln L}{\partial \kappa^2}\right) = \frac{1+\kappa}{1-\kappa} \cdot \sum_j \frac{n_j p_j q_j}{(p_j + q_j\kappa)(q_j + p_j\kappa)},$$

$$I_{0j} = -E\left(\frac{\partial^2 \ln L}{\partial \kappa \partial p_j}\right) = \frac{n_j(p_j - q_j)\kappa}{(p_j + q_j\kappa)(q_j + p_j\kappa)},$$

$$I_{jj} = -E\left(\frac{\partial^2 \ln L}{\partial p_j^2}\right) = \frac{n_j\{2p_jq_j(1-\kappa)(1-2\kappa) + \kappa(2-\kappa)\}}{p_jq_j(p_j + q_j\kappa)(q_j + p_j\kappa)}$$

for $j = 1, 2, \ldots, J$,

$I_{jj'} = 0$ for $j \neq j'$.

Note that the subscripts 0 and j of the information matrix are related to partials with respect to κ and p_j for notational simplification. Denote the vectors of parameters and scores as $\pi = (\kappa, \mathbf{p})$ and $\mathbf{S}(\pi) = (S_6(\pi), S_1(\kappa, p_1), \ldots, S_j(\kappa, p_j))$ where

$$S_\kappa(\pi) = \sum_j S_\kappa(\kappa, p_j) = \frac{1}{1-\kappa} \cdot \sum_j \left(\frac{x_{2j}}{p_j + q_j\kappa} + \frac{x_{0j}}{q_j + p_j\kappa} - n_j\right)$$

and

$$S_j(\kappa, p_j) = \frac{x_{2j} + x_{1j}}{p_j} - \frac{x_{0j} + x_{1j}}{q_j} + (1-\kappa) \cdot \left(\frac{x_{2j}}{p_j + q_j\kappa} - \frac{x_{0j}}{q_j + p_j\kappa}\right)$$

for $j = 1, 2, \ldots, J$. Letting the initial vector of π be $\tilde{\pi}^{(o)} = (\hat{\kappa}, \hat{p}_1, \ldots, \hat{p}_J)$, we have the 1st iterated values as

$$\tilde{\pi}^{(1)} = \tilde{\pi}^{(0)} + [I]^{-1}_{\pi=\tilde{\pi}^{(0)}} \cdot S\left(\tilde{\pi}^{(0)}\right).$$

The process is repeated until it converges.